

# The PathoSystems Resource Integration Center: Implications for Biodefense Viral Pathogens



## Abstract

The PathoSystems Resource Integration Center, PATRIC was established at VBI to create a public, web-based, comprehensive bioinformatics resource to the priority pathogens (the **bacteria**: *Brucella*, *Coxiella*, and *Rickettsia* and the **viruses**: Caliciviruses, Coronaviruses, Hepatitis A, Hepatitis E, and Lyssaviruses). These organisms are categorized by the National Institute for Allergy and Infectious Diseases (NIAID) as Category A, B, C organisms, depending on their potency and possibility of being used as bioterrorist weapons. The ultimate goal is to facilitate the research on the development of vaccines, diagnostics, and therapeutics. The data types covered by the resource include genome sequence, comparative genomics, polymorphisms, gene expression, proteomics, pathways and host/pathogen interactions. A new automated genome annotation pipeline, a new web-based curation interface, and a user-friendly web portal were developed. High-throughput and high quality annotation is facilitated by applying Reference Genome (RG) annotations in a controlled way to their Associated Genomes (AG). Curation means re-annotation of published genomes and annotation of novel genome sequences. Results are available at <http://patric.vbi.vt.edu> through query, visualization and analysis tools. PATRIC works closely with scientists in each organism community. PATRIC provides previously-available and updated annotations on all viral pathogens. Additional tools are in development to perform phylogenetic analysis and literature curation, the prototypes of which will be presented. A system to generate universal primers, useful as pathogen diagnostics, has been developed to design primers against rabies sequences and is being automated. We welcome and expect the research community to actively participate in the development of PATRIC as a useful computational resource for infectious disease research.

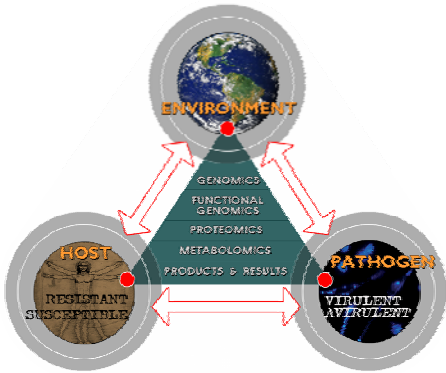


Figure 1. PathoSystems Biology involves interactive approach to infectious disease research that encompasses the environmental component.

## Introduction

In 2004, NIAID established VBI as one of the eight Bioinformatics Resource Centers (BRCs) with the mandate to annotate and (*manually*) curate genomic sequences of category A-C pathogens. Annotation is being done by identifying all genes, protein coding genes non-coding features. Providing protein characterization through structural features and functional attributes as well as pathway analysis that would show metabolic, signaling pathways and complex formation. At the same time, displaying comparative genomics intended to show phylogenetic analysis by comparing orthologous genes/proteins, hence revealing closely associated genomic structures and rearrangements. The study of these variations in the genome structure provides a useful avenue for comparative genomics, understanding host-pathogen interaction (Figure 1) and comparison of virulence in the genes responsible for disease. By using computational methods along with experimental evidence, it is possible to predict vaccine candidates, diagnostic methods and therapeutics (Figure 2). PATRIC project uses annotation and curation procedures necessary to improve the quality and quantity of data for the end user (Figure 3).

Viral genome projects have shown that different strains of the same species contain regions that are highly conserved across all strains as well as strain-specific, and non-conserved regions. The set of conserved regions is called the core genome, whereas the non-conserved, or variable genome regions are often called flexible genome. The PATRIC project-specific viruses are shown in Figure 4.

## Goals of PATRIC project

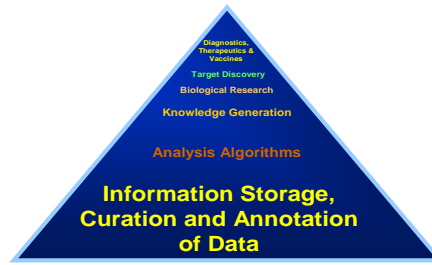


Figure 2. The goals of PATRIC are centered on data management and generation of applicable knowledge.

## Data available at [patric.vbi.vt.edu](http://patric.vbi.vt.edu) (some still under development)

All Proteins the genome codes for:  
Function of each protein  
Experimental characterization  
Cellular Localization  
Structure (Secondary and/or 3D when available)  
Active sites reported  
Pathways Involved/Affected (maybe host)  
Phylogenetic trees of different strains (from Literature)  
Host Pathogenicity  
Protein interactions (Protein-Protein, Protein-drug, Protein-DNA, Protein-RNA, Protein-Small molecule)  
Regulatory features on genome sequence  
Epidemiology  
Literature: Author, Journal, Publication, etc.  
Current Status of Drug/Vaccine/Diagnostics  
Reasons for absence of a vaccine/drawbacks

## Viral Curation Strategies

Complete or partial sequences from GenBank, and sequencing laboratories affiliated to PATRIC are uploaded into the database (Db) (Figure 4). A reference sequence or a set of sequences are chosen with the express agreement from an organism expert (OE). At the start of annotation, all the reference genomes (Figure 5) are fully curated followed by manual transfer of curated features to the associated genomes, use of statistical prediction tools (e.g GATU<sup>9</sup>) and literature also help in the annotation. PATRIC has been using a bottom-up curation method to clean-up and standardize all poorly curated genes from GenBank (Figure 2), but a more efficient top-down curation or targeted annotation will help in quickly identifying genes of interest for generation of experimental data. Other curation strategies include complete data integration by combining diverse exchange(s) of data types in the annotation process through pipelines.

## PATRIC Annotation Workflow

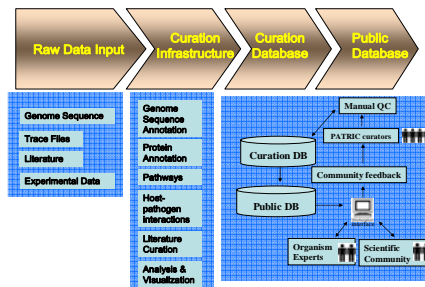


Figure 3. Annotation workflow used by PATRIC team to produce data and information for end users.

PATRIC Organisms (Viruses)	Named Isolates	Complete Genomes	Genome Length in kilobases (kb)	ORFs	Associated Diseases
Caliciviruses	1832	72	8	3	Food-borne Gastroenteritis
Coronaviruses	286	184	30	14	SARS
Hepatitis A viruses	14	16	7	1	Hepatitis
Hepatitis E viruses	9	48	7.5	3	Hepatitis
Lyssaviruses	13	12	12	5	Rabies

Figure 4. Viruses being annotated by PATRIC, showing the total number of complete genomes, length (kb), number of open reading frames (ORFs) per strain, and diseases associated with each virus group.

## Pathosystem

## No. of Reference Genomes

Caliciviridae	13
Coronavirus	16
Hepatitis A virus	6
Hepatitis E virus	5
Lyssavirus	1

Figure 5. The total number of reference genomes for each virus group being curated by PATRIC.

## Viral Curation and Annotation Activities by PATRIC

- All viral reference genomes have been annotated at nucleic acid (NA) level.
- Phylogenetic trees: Generated for Calci-, Hep-A, Hep-E and Lyssaviruses using ORF1, VP1, whole genome sequences and nucleoproteins, respectively.
- Standard operating procedures (SOPs): SOPs are developed for all viral curations and provide standardized methods across all annotations.

### Caliciviridae (Calci-)

Associated genomes in the four genera of Caliciviridae are being annotated; there is little data on the size and nature of the mature peptides. PATRIC is using computational (GATU) and statistical (Gen-Var<sup>8</sup>) approaches to identify the mature peptides in all the genome sequences. This approach will allow us to standardize *Calicivirus* annotations.

### Coronavirus (Corona-)

A portion of the associated genomes have been annotated using features from the reference genomes. These include identification of all the known genes, coding and mature peptides and standardization names for all features. Other non-coding features are the leader sequence; transcriptional regulatory sequences; 5' and 3' un-translated (UTR) regions; RNA pseudoknot; RNA slippery sequence.

### Hepatitis A viruses (Hep-A)

Coding features from the reference genome have been transferred using GATU. In addition to standardizing feature annotation and nomenclature, non-coding features such as internal ribosome entry site (IRES), secondary structural domains I-VI (SSD I-VI) and polypyrimidine tracts have been added from the literature.

### Hepatitis E viruses (Hep-E)

This manual annotation has standardized the names of the open reading frames, corrected coordinates based on the current literature. The 5' 3' UTRs and polypyrimidine tracts have been added. Coding and non-coding features have been annotated from literature and standardized across all strains, including alternate translational start sites and potential overlapping features.

### Lyssaviruses (Lyssa-)

Standardized names have been inserted in the associated genomes for the genes/proteins and the annotations of non-coding features such as leader RNA, alternate in-frame start codons, polyadenylation sites, and consensus mRNA start sequences. For the summary of PATRIC curated features on viruses (see Figure 6).

## Future Trends at PATRIC

In the next six months (June to December 2006) PATRIC will avail:

- A Phylotyping tool for use on viruses - (in collaboration with the Center for Disease Control and Prevention, CDC)
- A Universal Primer Design tool for use across the viral genomes - (CDC)
- Epitope Predictions and tools in collaboration with Immune Epitope Database and Analysis Resource (IEDB)

Curated Feature	Calci-	Corona-	Hep-A	Hep-E	Lyssa-
CDs	181	413	30	165	64
RNA	nc	nc	nc	nc	10
Gene	161	295	30	165	64
mRNA	nc	nc	nc	nc	5
Mature Peptide	67	540	184	nc	1
Misc. Feature	nc	654	47	nc	161
Misc. difference	nc	2	nc	nc	nc
Misc. Structure	nc	nc	58	nc	nc
5'UTR	17	165	nc	38	nc
3'UTR	18	3	nc	27	nc
Poly (A) site	72	nc	nc	4	nc

Figure 6. A number of features on viral genomes have been curated by the PATRIC team (nc=not curated).

## Summary

The PATRIC project at VBI is currently providing high quality data and information as well as developing analysis tools to research community studying Caliciviruses, Coronaviruses, Hepatitis A, Hepatitis E, and Lyssaviruses. The integration of manual genome/gene product curation, scientific literature, and a suite of comparative genomic tools will align with NIAID's mission to facilitate development of vaccines, diagnostics and therapeutics. The project welcomes all scientists to utilize its freely available resources at the PATRIC website.

## References

- <sup>8</sup>Tcherepanov VT, Ehlers A, Upton C. (2006) Genome Annotation Transfer Utility (GATU): Rapid annotation of viral genomes using a closely related reference genome. BMC Genomics Jun 13:711-150.
- <sup>9</sup>Yu G. X., M. Czar, A. Purkayastha, E.E. Snyder, S. M. Boyle, O.R. Crasta, J. Setubal, B. Sobral. (In preparation). Gen-Var, a Genome-context-based Procedure for Critical Analysis of Sequence Variants: *Brucella* as a Model Organism.

## Acknowledgement:

This project is funded by NIAID/National Institute of Health (NIH) contract HHSN26620040035C awarded to B.W. Sobral. Our thanks to Drs. Jan Vinjé, Susan Baker, Yuri Khudiyakov, X. J. Meng, and Charles Rupprecht OEs for Caliciviruses, Coronaviruses, Hepatitis A, Hepatitis E & Lyssaviruses, respectively, and to June Mullins for preparing the poster.